



# Haplotype analysis in population genetics and association studies

Hongyu Zhao<sup>†1,2</sup>,  
Ruth Pfeiffer<sup>2</sup> &  
Mitchell H Gail<sup>2</sup>

<sup>†</sup>Author for correspondence  
<sup>1</sup>Department of Epidemiology  
and Public Health, Yale  
University School of  
Medicine, New Haven,  
CT 06520, USA  
<sup>2</sup>Biostatistics Branch,  
Division of Cancer  
Epidemiology and Genetics,  
National Cancer Institute,  
Bethesda, MD 20892, USA  
Tel: +1 203 785 6912;  
Fax: +1 203 785 6271;  
E-mail: hongyu.zhao@  
yale.edu

Several studies of haplotype structures in the human genome in various populations have been published recently. Such knowledge may provide valuable information on human evolutionary history and lead to the development of more efficient strategies to identify genetic variants that increase susceptibility to human diseases. In this review, we summarize the current understanding of haplotype structure, diversity, and distribution in the human genome, with a focus on statistical issues in using haplotypes for studies of population genetics and evolutionary history, as well as to identify genetic variants underlying complex human traits.

## Introduction

The Human Genome Project and other large-scale efforts have identified millions of genetic markers that can be used in genetic studies. Although each marker can be analyzed independently of other markers, it is much more informative to analyze markers in a region of interest simultaneously. The combination of marker alleles on a single chromosome is called a haplotype (*Haplôid Genotype*). There is great interest in understanding haplotype structures in the human genome using identified genetic markers because: 1) haplotype structures may provide critical information on human evolutionary history and the identification of genetic variants underlying various human traits; and 2) molecular technologies now make it possible to study hundreds of thousands of genetic polymorphisms in population samples of reasonable sizes. For haplotypes including markers tightly linked with each other, for example, markers within the same gene, alleles at these markers often display statistical dependence, a phenomenon called linkage disequilibrium (LD), or allelic association. One major aspect of haplotype analysis is to identify LD patterns in different regions and different populations because the very existence of LD among markers makes it possible to infer population histories and localize genetic variants underlying complex traits. It is well known that LD is affected by many factors, including the age of the variants, population history, recombination rates, gene conversion, natural selection, and other factors. Although LD can be studied through theoretical population genetics models, many recent empirical studies have shown that available theoretical models are not

able to explain the complex haplotype structures in the human genome. We review the haplotype structures revealed by these empirical studies in general populations and provide an overview of statistical methods that have been useful in analyzing haplotypes. In addition, we discuss statistical methods to identify candidate regions associated with complex traits using haplotypes, and we review various methods that have been proposed to reduce the dimensionality of haplotype analysis. We conclude this review by pointing to several directions that need vigorous developments in the next several years to fully realize the potential in haplotypes.

Population genetics of haplotypes and  
linkage disequilibrium  
*Linkage disequilibrium*

A study of haplotypes consisting of a short tandem repeat polymorphism (STRP) and an Alu deletion polymorphism at the CD4 locus in 42 worldwide populations first demonstrated the usefulness of comparing LD patterns of different populations for inferring population history [1]. The LD patterns between these two polymorphisms provided evidence of a common and recent African origin for all non-African populations. This was the first study using autosomal regions to provide evidence on the out-of-Africa hypothesis, in contrast to studies using mitochondrial DNA and Y chromosomes. Long-range haplotype structure may also provide a more powerful tool to detect recent selection in the human genome [2].

In addition to its use in inferring population history, the extent of LD is a critical factor in identifying disease-associated genetic variants

**Keywords:** complex disease,  
genetic association, haplotype,  
linkage disequilibrium,  
population genetics



Ashley Publications Ltd  
www.ashley-pub.com

and designing efficient studies to detect disease-gene associations. There are many measures to quantify the degree of association between two polymorphisms [3]. The most commonly used ones are  $D'$  and  $r^2$ . Both  $|D'|$  and  $r^2$  range between 0 and 1. Simulation studies based on simple population genetics models suggested that useful LD extends only a few kilobases (kb) around common single nucleotide polymorphisms (SNPs) [4]. However, empirical data imply that LD can extend much further. For example, a systematic study of 19 chromosomal regions found that  $|D'|$  drops below 0.5, on average at ~ 60 kb in a Utah population and on average at ~ 5 kb in Yoruban samples [5]. It is also important to choose appropriate LD measures to characterize LD patterns. For example, the use of  $|D'|$  as a measure was criticized as it is biased upwards inversely with sample size [6]. Therefore, different sample sizes from different populations may confound LD comparisons. A systematic study of markers on chromosome 22 using European samples showed similar patterns [7]. The discrepancy between model-based simulation results and empirical observations suggests that the oversimplifying assumptions in the original simulation study did not adequately account for important factors that determine LD. Alternative simulations based on more realistic population genetics models that include population bottlenecks and non-uniform recombination rates demonstrated that population bottlenecks were not adequate in explaining the observed LD patterns, mostly because these models cannot simultaneously accommodate the observed extent of LD and the level of genetic polymorphism in the genome [8].

Two major features have emerged from many studies on the extent of LD at different loci and in different populations. First, although LD between two markers tends to decrease as their physical distance increases, the variation is so great that it is not possible to predict LD between two polymorphisms reliably, based only on their physical distance. Second, the amount of LD differs among different populations, and LD is usually weaker among Africans than other populations [9-11]. This pattern of variation strongly supports the hypothesis of recent expansion of the human population. In summary, LD is both locus and population specific, and it is impossible to understand LD patterns in the human genome without a systematic empirical study that covers the genome and involves many human populations. The fact that different

populations display different LD suggests that an initial genome-wide association study may be conducted in a population with stronger LD, and the variants can be fine mapped in populations with less LD [12].

The magnitude of LD in a local region tends to be small if the estimated local recombination rate is high [5,7]. Since the recombination rate is associated with many characteristics of the DNA sequences [13], it is not surprising that LD is also significantly associated with sequence properties [7].

#### *Haplotype blocks*

Some recent studies have found that the chromosomes are structured such that each chromosome can be divided into many blocks, within which there is limited haplotype diversity. Although block structures were found in these studies, there is no universally accepted definition of haplotype blocks. In fact, each study had its own definition. Some examples of definitions of haplotype blocks are:

- a contiguous set of markers in which the average  $D'$  is greater than some predetermined threshold [5]
- a region where a small number of common haplotypes account for the majority of chromosomes [14,15]
- a chromosomal segment with reduced levels of haplotype diversity
- regions with both limited haplotype diversity and strong LD except for a few markers [7]
- regions with absolutely no evidence for historical recombination between any pair of SNPs [16]

The block structures identified in each study depend strongly on the definition used, and there has been no systematic comparison of haplotype blocks identified under these various definitions. Which definition is most appropriate may depend on how the inferred blocks will be used. Different goals may be, for example, to infer recombination hot spots or to identify regions that are associated with complex traits.

In the following, we summarize some results on haplotype blocks but stress that the identified blocks were based on various definitions and their biological relevance still needs to be demonstrated. When a 500 kb region on chromosome 5 was studied using 103 common SNPs in 129 European trios, discrete haplotype blocks with limited diversity punctuated by apparent sites of recombination were found; the size of these blocks ranged up to 100 kb [17]. A study of

51 autosomal regions in four populations found that the minimal span of the blocks averaged 9 kb in Yoruban and African-American samples with a range of < 1 kb up to 94 kb, whereas the average in European and Asian samples was 18 kb with a range of < 1 kb to 173 kb [11]. Block structures were also found in systematic studies of chromosome 21 [14] and chromosome 22 [7].

Although haplotype block boundaries have been found to correlate with recombination hot spots [18], haplotype blocks can arise without recombination hotspots due to other factors that affect LD patterns, such as genetic drift [16,19]. More data are needed to evaluate the usefulness of haplotype blocks as a general tool to identify recombination hot spots. Although there are usually recombination events between blocks, LD is often found between loci in different blocks, and this LD can be substantial [11]. Therefore, using each block as a unit to study association between complex traits and candidate regions may not be the most efficient strategy.

Estimates of haplotype frequency and of haplotype pairs carried by an individual Although it is possible to infer two haplotypes from each individual through molecular methods [14], such methods are currently too expensive and laborious to be practical in large-scale population studies. Information from relatives may help resolve haplotype ambiguity, but such ambiguity may still exist even with data from many relatives, especially as the number of markers increases. Therefore, several assumption-based numerical methods have been developed to infer haplotypes and estimate haplotype frequencies. The central Hardy-Weinberg assumption underlying these methods is that each haplotype carried by an individual represents an independent sample from the population of haplotypes.

Clark first proposed an algorithm to infer haplotypes among unrelated individuals [20]: haplotypes are determined from those individuals with no haplotype ambiguity, and then ambiguous individuals are considered sequentially to resolve their haplotypes. Several groups applied the Expectation-Maximization (EM) algorithm [21] to obtain the maximum likelihood estimates of haplotype frequencies in the sample [22-24]. The EM algorithm uses an initial set of haplotype frequency estimates to calculate the conditional distribution for haplotype pairs that each individual carries (Estimation step). Based on these conditional distributions, haplotype frequency estimates can be updated

(Maximization-step). The EM algorithm iterates between these two steps until haplotype frequency estimates converge. Despite its simplicity, standard applications of the EM algorithm may not be feasible when analyzing many markers simultaneously, as the number of haplotypes that needs to be considered increases exponentially with the number of markers. Several algorithms try to circumvent this limitation [25]. The PHASE program [26] and its modified version [27] use coalescent models and Markov chain Monte Carlo methods to assign phases in each individual and estimate haplotype frequencies. These algorithms use population genetics models to relate different haplotype patterns such that a haplotype that is more similar to the commonly observed haplotype patterns is more likely to be inferred to be present than less similar haplotypes. Several other Bayesian methods and modified EM algorithms have also been developed to facilitate haplotype analysis for many markers simultaneously [28]. Although some comparisons of different methods have been made in the above mentioned studies, it is likely that no method is uniformly best. There remains a need for further comparative studies and new methodology. In addition, haplotype inference in (large) families remains a challenging problem.

In haplotype inference, one crucial, yet under appreciated, step is choosing the set of markers to be analyzed. Although some methods are able to handle many markers, there may still be too many possible haplotypes to estimate frequencies, especially if there is weak LD among them. In the most extreme case, if all the markers to be analyzed are in linkage equilibrium, the number of possible haplotypes may be too large for available algorithms. On the other hand, although phase inference is needed in a region with many tightly linked markers, it may not be necessary to infer haplotype for all markers in a long region, say over a 10 cM interval, if the focus is on identifying disease variants in a local region, usually within a candidate gene.

Although many algorithms produce the distribution of haplotype pairs for an individual, some programs only produce the most likely haplotype pair. This practice may lead to loss of information and potential bias in further analyses that regard inferred phases as observed phases. For example, consider an individual who has a 55% probability of carrying one haplotype pair and a 45% probability of carrying another haplotype pair. If only the more likely pair are kept and used in association analyses, information about

the other pair will be lost and this may lead to both loss of efficiency and bias in assessing association between this region and traits of interest.

With appropriate statistical tools, haplotype frequencies can be estimated from directly observed haplotypes, from diploid individuals where phases may be ambiguous [29], from related individuals [30,31], and from DNA pooling with two or more individuals [32,33]. Although for a fixed sample size, molecular haplotyping methods can produce much more precise estimates of haplotype frequencies than other approaches, the considerable cost in obtaining and analyzing individual chromosomes may make alternative designs preferable [30]. Empirical data also indicate advantages from using family data, including detection of genotyping errors and integration with meiotic maps [7].

#### Association analysis with haplotypes

When multiple markers, often in LD, in a chromosomal region are studied to assess the association between this region and traits of interest, a statistical analysis based on haplotypes may be more efficient than separate analyses of the individual markers. This has been demonstrated both through simulation studies [34] and empirical studies [35,36].

Statistical methods that use haplotypes to map disease genes can be broadly divided into two categories: those developed to locate the exact chromosomal location (i.e., the exact base pair) of disease-susceptibility variants, and those developed to locate a general chromosomal region with disease (i.e., a candidate gene region). For the first class of methods, the underlying principle is that more similarity among the chromosomes of diseased subjects should be observed near the loci of genetic variants that increase disease susceptibility. This general principle can be realized via various test statistics. One approach is to explicitly model the origin of the disease variant(s), their evolution in the general population, as well as the history of the general population. Such methods are model-based, and as a class, are called LD mapping methods [37]. Another class of methods does not explicitly model the evolution of haplotypes carrying disease susceptibility variants but uses test statistics based on heuristic reasoning. For example, the Haplotype Sharing Statistic [38] scans the candidate region and compares the mean length of haplotype sharing on chromosomes of cases with disease versus chromosomes of control subjects without disease. For these fine-mapping methods,

the candidate position is varied systematically in a region and a score defining the evidence for genetic association is calculated. The position where the largest score is obtained is usually taken as the most likely site for a disease susceptibility variant. Although such methods have successfully mapped genetic variants for Mendelian diseases in isolated populations, their utility for identifying common disease-susceptibility variants that predispose to complex diseases still needs to be demonstrated empirically. In fact, when the primary focus is a candidate region of limited length, i.e., several hundred kilobases, strong dependence among markers and a lack of clear relationship between physical proximity and LD within a local region may make fine mapping difficult or impossible. For example, when markers are in perfect LD, there is no statistical means to distinguish which marker is the true functional one. It is likely that the final identification and understanding of genetic variants responsible for disease susceptibility will rely on ancillary experiments to define the functional effects of the variants. Therefore, a more realistic and achievable objective for haplotype analysis is to localize a region that contains disease susceptibility variants.

To assess associations between haplotypes in a candidate region and traits of interest, the simplest models regress the trait on the two haplotypes each individual carries as well as on other factors, such as age, gender and smoking status. The trait might be continuous, such as blood pressure, or discrete, such as the presence or absence of disease, or the time to disease onset. If there is no association between this region and traits of interest (null hypothesis), all the haplotypes are expected to have the same effect on trait values. However, different haplotypes should affect phenotypes differently when an association exists. Therefore, under the alternative hypothesis, we allow different haplotypes to affect traits differently, and interactions among haplotypes as well as interactions between haplotypes and environmental factors may be taken into account. Standard statistical methods in linear regression analysis can be used to investigate association between continuous traits and haplotypes, whereas logistic models [39] can be used to analyze binary traits, and survival models can be used for time to disease-onset data.

Two major issues complicate these simple regression analyses in samples of unrelated individuals. The first issue is haplotype uncertainty. As discussed above, haplotypes may have

to be inferred from marker phenotype data unless molecular haplotyping methods are used. Therefore, haplotype uncertainty needs to be incorporated into the regression approach to analyzing associations between haplotypes and disease outcomes. The second issue is haplotype complexity. Although haplotypes may be more informative than single markers, the power of haplotype analysis is reduced by the potentially large number of haplotypes that needs to be studied. Statistical approaches that have been developed to address these two issues are outlined below.

#### *Statistical approaches to address haplotype uncertainty*

In the presence of haplotype uncertainty, an individual can be assigned to have different haplotype pairs with different probabilities. These probabilities depend on the specific method used for phase inference. Then statistical models that directly model an individual's phenotype as a function of each inferred haplotype pair, weighted by their estimated probability, can be used for association studies. This approach has led to a mixture model for association between continuous traits and haplotypes [40], and general score tests between binary and continuous traits and haplotypes [41]. Statistical methods that simply compare haplotype frequencies between cases and controls without explicitly modeling trait-haplotype associations have also been proposed [42].

#### *Approaches to deal with haplotype complexity*

Two classes of methods have been developed to reduce the number of haplotypes considered in association studies. The first class of methods divides the whole chromosomal region into smaller regions for analysis, whereas the second class groups haplotypes into a smaller number before association analysis. The first class of methods generally has a sliding window on the candidate region and assesses evidence for association within each window [38,43-45]. The sliding window serves two purposes. First, the number of haplotype patterns in each window may be significantly less than that in the whole region, so the regression analysis involves fewer parameters and likely has better power if there is an association between disease and haplotypes. Second, it is anticipated that evidence near the true disease variants is stronger than that in other regions. Therefore, these methods resemble fine disease-mapping methods.

The second class of methods has a somewhat longer history [46-49]. The central assumption is that an unknown mutation causing a phenotypic effect occurred at some point in the evolutionary history of the population and became embedded within the historical structure represented in a tree structure relating different haplotypes, called a cladogram. Evolutionary principles suggest that certain portions of the cladogram would display the phenotypic effect while other portions would not. Thus, the cladogram defines a nested analysis of variance that simultaneously detects phenotypic effects and localizes the effects within the cladogram.

It has been argued that haplotype block structures can be helpful for association studies because each haplotype block can be treated as a locus with several alleles (the block-specific haplotypes) in association studies [17]. However, the results depend on the definition of haplotype blocks, and such methods may not be the most efficient ones if there is substantial LD among alleles in different blocks [11].

#### *Conclusion and outlook*

There is a growing belief that haplotypes may hold the key to better understand human evolutionary history and to more efficiently identify genetic variants underlying complex traits. Although great progress has been made and much has been learned in the past several years about haplotype structures in the human genome, much work remains to adequately characterize haplotype and LD patterns throughout the genome because such patterns are both locus and population specific and theoretical models have limited predictive value. Even more challenging than the need for empirical data is the difficulty in synthesizing and interpreting haplotype data to learn what factors shape haplotype structure and how to use haplotypes to assist in localizing disease susceptibility variants. There is also an urgent need to develop efficient designs to conduct genetic association studies and to develop statistical tools to analyze data from these studies. Although there is a general belief that haplotype analysis might be more powerful than single marker analysis, the true power of haplotype analysis still needs to be demonstrated both in theoretical studies and in practice. Statistical methods need to incorporate knowledge of haplotype patterns, and statistical findings need to be interpreted in combination with prior biological information that can be extracted with bioinformatics tools.

Although statistical methods have been developed for haplotype associations using unrelated individuals, pedigree data are routinely collected in genetic studies. Haplotype inference and analysis in general pedigrees is an area that needs vigorous development in the next several years to realize the information in pedigrees fully. In addition, statistical methods are needed to appropriately analyze more complex data types: for example, survival data, ordinal data, and longitudinal data.

In genome-wide association studies, tens of thousands of regions are studied to detect possible associations with traits of interest, and the statistical power to identify a true association may be drastically reduced by the need to adjust for the large number of statistical tests to be performed. Even within a single region, the multiple comparison issue may arise because of a potentially large number of haplotypes present in that region. Although recent developments in multiple comparison adjustments, such as the concept

of false discovery rate [50] instead of the family-wise type-I error rate, may lead to better strategies to identify true associations, statistical methods that incorporate prior biological information into the data analysis may improve the power to detect true associations. In addition, better procedures are needed to reduce the dimensionality of the space spanned by haplotypes in a local region in order to ameliorate power loss from multiple comparisons.

Although millions of markers are available, LD among markers in a local region can be used to reduce the number of the markers that need to be studied. Several methods have been proposed to select representative markers based on a sample from the general population [51-53]. One use of these representative markers is to study various aspects of population genetics. For this purpose, selecting markers that capture the diversity or otherwise represent the full set of markers seems reasonable, and selection can be based on a sample of unaffected individuals. A second goal is to select a set of markers to detect an association of disease with haplotype. For this purpose, the strategies for selecting subsets of markers based on diversity or representative are not specifically tailored for designing powerful studies of genetic associations with disease. These strategies may lead to poor choices of markers for detecting disease associations. For example, if we knew a disease susceptible variant, there would be no need to study other markers near this variant. In the case that this variant is rare and we select markers based on diversity, it is likely that we would choose to study more polymorphic markers that are not as strongly associated with the disease of interest as this 'rare' variant, leading to reduced power to identify disease-marker associations. Alternative strategies for marker selection may prove to be more powerful and appropriate in disease association studies.

Although it is impossible to distinguish a true disease susceptibility genetic variant from a marker that is in perfect LD with this variant using association studies, it may be possible if the LD is not perfect [54]. However, a very large sample may be needed to distinguish these two markers to identify the truly functional one. It is not clear whether it is feasible to identify functional variants statistically in light of the LD patterns emerging from recent empirical studies.

Another issue that we have not addressed is the potential bias in association studies resulting from population stratification. Many methods [55-57] have been developed to use a large number

## Highlights

- Haplotype structures may provide critical information on human evolutionary history and the identification of genetic variants underlying various human traits through linkage disequilibrium (LD), or allelic association. LD patterns in different regions and different populations make it possible to infer population histories and localize genetic variants underlying complex traits.
- LD is affected by many factors, including the age of the variants, population history, recombination rates, gene conversion, natural selection, and other factors. The relative contributions of these factors to LD patterns in human populations are an active area of research.
- Empirical studies suggest that the chromosomes are structured such that each chromosome can be divided into many blocks, within which there is limited haplotype diversity. However, the degree of LD is both locus specific and population specific, with LD usually weaker among Africans than other populations.
- Several computational methods have been developed to infer individual haplotypes from marker phenotypes. Although it is likely that no method is uniformly best, more comparative studies and new methodology are needed for haplotype inference.
- When multiple markers in a chromosomal region are studied to assess the association between this region and a complex disease, a statistical analysis based on haplotypes may be more efficient than separate analyses of the individual markers. However, the true power of haplotype analysis still needs to be demonstrated both in theoretical studies and in practice.
- The most important issue in haplotype analysis is to incorporate knowledge of haplotype patterns and reduce the number of haplotypes considered. Statistical findings also need to be interpreted in combination with prior biological information that can be extracted with bioinformatics tools. Both these areas need methodology developments.
- Other issues in haplotype analysis that need theoretical and empirical investigations include genetic marker selection, multiple comparisons, population heterogeneity, and study designs.

of markers either to identify the underlying population structure or to adjust for bias due to population stratification. Such methods can be applied to haplotype analysis to make the results robust to population stratification.

Finally, it is still not clear what types of genetic variants, such as common alleles or rare alleles, underlie most common diseases [58,59]. The exact nature of these variants shall influence our strategy to identify them, and future developments in designs and analytical methods for association studies will depend on our increasing knowledge of the biology of disease-inducing variants. Although haplotype analysis may increase our power to map disease genes

compared to single marker analysis, it is possible that heterogeneity among disease variants resulting in the same phenotype may seriously limit our ability to identify each individual variant, even with the help of haplotypes. The choice of study population and sampling design, and the incorporation of epidemiologic and genetic information will be important considerations in planning and executing an association study to identify disease variants.

#### Acknowledgement

We thank two anonymous reviewers for their constructive comments. Supported in part by grant GM59507 from the US National Institutes of Health.

#### Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Tishkoff, SA, Dietzsch E, Speed W *et al.*: Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271, 1380-1387 (1996).
- The first study demonstrating the use of LD in inferring population history.
2. Sabeti PC, Reich DE, Higgins JM *et al.*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832-837 (2002).
3. Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311-322 (1995).
4. Kruglyak L: Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139-144 (1999).
5. Reich DE, Cargill M, Bolk S *et al.*: Linkage disequilibrium in the human genome. *Nature* 411, 199-204 (2001).
- Systematic study of LD in 19 chromosomal regions in a north-European descent population and a Nigerian population.
6. Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18, 19-24 (2002).
7. Dawson E, Abecasis GR, Bumpstead S *et al.*: A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418, 544-548 (2002).
8. Reich DE, Schaffner SF, Daly MJ *et al.*: Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* 32, 135-142 (2002).
9. Stephens JC, Schneider JA, Tanguay DA *et al.*: Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293, 489-493 (2001).
- An LD study on 313 genes in 82 unrelated individuals of diverse ancestry.
10. Pakstis AJ, Zhao H, Kidd JR, Kidd KK: Patterns of linkage disequilibrium for multisite haplotypes at 14 loci in human populations worldwide. *Am. J. Hum. Genet.* 67, A84 (2000).
11. Gabriel SB, Schaffner SF, Nguyen H *et al.*: The structure of haplotype blocks in the human genome. *Science* 296, 2225-2229 (2002).
12. Risch NJ: Searching for genetic determinants in the new millennium. *Nature* 405, 847-856 (2000).
13. Kong A, Gudbjartsson DF, Sainz J *et al.*: A high-resolution recombination map of the human genome. *Nat. Genet.* 31, 241-247 (2002).
14. Patil N, Berno AJ, Hinds DA *et al.*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294, 1719-1723 (2001).
- The first systematic chromosome-wide study of LD.
15. Zhang K, Deng MH, Chen T, Waterman MS, Sun FZ: A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. USA* 99, 7335-7339 (2002).
16. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L: Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* 71, 1227-1234 (2002).
17. Daly MJ, Rioux JD, Schaffner SE, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229-232 (2001).
- The authors show the presence of discrete haplotype blocks in a 500 kb region on chromosome 5q31 in a European-derived population.
18. Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29, 217-222 (2001).
19. Subrahmanyam L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA: Sequence variation and linkage disequilibrium in the human T-cell receptor  $\beta$  (TCRB) locus. *Am. J. Hum. Genet.* 69, 381-395 (2001).
20. Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* 7, 111-122 (1990).
21. Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* 39, 1-38 (1977).
22. Excoffier L, Slatkin M: Maximization-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12, 921-927 (1995).
23. Hawley M, Kidd KK: HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.* 86, 409-411 (1995).
24. Long J, Williams R, Urbanek M: An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* 56, 799-810 (1995).
25. Qin ZHS, Niu TH, Liu JS: Partition-ligation-expectation-maximization

- algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 71, 1242-1247 (2002).
- **The authors develop variations of the EM algorithm for haplotype inference.**
26. Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978-989 (2001).
- **The authors propose statistical methods based on coalescent models for haplotype inference.**
27. Lin S, Cutler DJ, Zwick ME, Chakravarti A: Haplotype inference in random population samples. *Am. J. Hum. Genet.* 71, 1129-1137 (2002).
28. Niu TH, Qin ZHS, Xu XP, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 70, 157-169 (2002).
29. Douglas JA, Boehnke M, Gillanders E, Trent JA, Gruber SB: Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat. Genet.* 28, 361-364 (2001).
30. Schaid DJ: Relative efficiency of ambiguous versus directly measured haplotype frequencies. *Genet. Epidemiol.* 23, 426-443 (2002).
- **A careful evaluation of the relative efficiency of various strategies to estimate haplotype frequencies.**
31. Becker T, Knapp M: Efficiency of haplotype frequency estimation when nuclear family information is included. *Hum. Hered.* 54, 45-53 (2002).
32. Pfeiffer RM, Rutter JL, Gail MH, Struwing J, Gastwirth JL: Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet. Epidemiol.* 22, 94-102 (2002).
33. Wang S, Kidd KK, Zhao H: On the use of DNA pooling to estimate haplotype frequencies. *Genet. Epidemiol.* 24, 74-82 (2003).
34. Morris RW, Kaplan NL: On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet. Epidemiol.* 23, 221-233 (2002).
35. Martin ER, Lai EH, Gilbert JR *et al.*: SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer's disease. *Am. J. Hum. Genet.* 67, 383-94 (2000).
36. Drysdale CM, McGraw DW, Stack CB *et al.*: Complex promoter and coding region  $\beta(2)$ -adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl. Acad. Sci. USA* 97, 10483-10488 (2000).
37. Lazzaroni LC: A chronology of fine-scale gene mapping by linkage disequilibrium. *Stat. Methods Med. Res.* 10, 57-76 (2001).
38. Van der Meulen MA, Te Meerman GJ: Association and haplotype sharing due to identity by descent, with an application to genetic mapping. In: *Genetic Mapping of Disease Genes*. Pawlowitzki I-H, Edwards JH, Thompson E (Eds), Academic Press, London 115-135 (1997).
39. Wallenstein S, Hodge SE, Weston A: Logistic regression model for analyzing extended haplotype data. *Genet. Epidemiol.* 15, 173-181 (1998).
40. Keavney B, McKenzie C, Connell JMC *et al.*: Measured haplotype analysis of the angiotensin-1 converting enzyme (ACE) gene. *Hum. Mol. Genet.* 7, 1745-1751 (1998).
41. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* 70, 425-434 (2002).
- **The authors develop statistical methods that can appropriately take haplotype uncertainty into account in disease-gene association studies.**
42. Fallin D, Cohen A, Essioux L *et al.*: Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res.* 11, 143-151 (2001).
43. Clayton D, Jones H: Transmission/disequilibrium tests for extended marker haplotypes. *Am. J. Hum. Genet.* 65, 1161-1169 (1999).
44. Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F: Search for multifactorial disease susceptibility genes in founder populations. *Ann. Hum. Genet.* 64, 255-265 (2000).
45. Toivonen HT, Onkamo P, Vasko K *et al.*: Data mining applied to linkage disequilibrium mapping. *Am. J. Hum. Genet.* 67, 133-145 (2000).
46. Templeton AR, Boerwinkle E, Sing CF: A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics* 117, 343-351 (1987).
47. Templeton AR, Sing CF, Kessling A, Humphries S: A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics* 120, 1145-1154 (1988).
48. Seltman H, Roeder K, Devlin B: Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am. J. Hum. Genet.* 68, 1250-1263 (2001).
49. Hoehe MR, Kopke K, Wendel B *et al.*: Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. *Hum. Mol. Genet.* 9, 2895-2908 (2000).
50. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* 57, 289-300 (1995).
51. Judson R, Salisbury B, Schneider J, Windemuth A, Stephens JC: How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* 3, 379-391 (2002).
52. Johnson GCL, Esposito L, Barratt BJ *et al.*: Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29, 233-237 (2001).
53. Stram DO, Haiman CA, Kolonel LN, Henderson BE, Pike MC: Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects: the multiethnic cohort study. Unpublished manuscript (2002).
54. Valdes AM, Thomson G: Detecting disease-predisposing variants: the haplotype method. *Am. J. Hum. Genet.* 60, 703-716 (1997).
55. Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 55, 997-1004 (1999).
56. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170-181 (2000).
57. Zhang S, Zhu X, Zhao H: On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet. Epidemiol.* 24, 44-56 (2003).
58. Pritchard JK: Are rare variants responsible for susceptibility to common diseases? *Am. J. Hum. Genet.* 69, 124-137 (2001).
59. Reich DE, Lander ES: On the allelic spectrum of human disease. *Trends Genet.* 17, 502-510 (2001).